UNIVERSITÉ DE MONTRÉAL

CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

ESTIMATING THE O-D MATRIX

FROM OBSERVED LINK VOLUMES

by

Pierre Robillard

This research was supported by a grant from the Ford Motor Company of Canada for applied research in the Analysis and Control of Urban Transportation Networks.

Université de Montréal
Publication # 152 - Département d'informatique
Publication # 1 - Centre de Recherche sur les Transports
Février 1974
Révisé août 1974

Dépôt légal - 4e trimestre 1978

Abstract

Until now the traffic volume data have been used almost exclusively for the study of circulation control and road construction. We present a methodology to use these data for transportation planning where traditionally the data needed came from an O-D survey. We demonstrate that the traffic volume data are informative enough to be useful for this puspose.

Most specifically we want to study the estimate of the O-D matrix obtained using observed link volumes for the class of proportional assignment methods. These methods are essentially the ones which do not take into account explicitly or implicitly any constraints of capacities.

We show that for this class of method the estimation of the O-D matrix is equivalent to a linear regression problem. We discuss the problem of estimability of the parameters and relate this question with the sample of traffic volume needed. The discussion is illustrated by an example.

Résumé

Jusqu'à date les données de volume de trafic ont été utilisées presqu'exclusivement pour étudier le contrôle de la circulation ou l'établissement de routes nouvelles. Nous présentons ici une méthodologie permettant l'utilisation de ces données pour la planification en transport où traditionnellement les données utilisées étaient celles d'une matrice O-D. Nous démontrons que les données de volume de trafic peuvent servir à la planification directement.

Plus spécifiquement nous voulons estimer les entrées de la matrice O-D à partir des volumes de trafic observés et ce pour la classe des modèles d'affectation proportionnelle. Ceux-ci sont ceux qui ne tiennent pas compte implicitement ou explicitement des contraintes de capacités.

Nous montrons que pour cette classe de modèle, l'estimation de la matrice O-D est équivalente à résoudre un problème de regression linéaire. Nous discutons en plus des problèmes d'estimabilités des paramètres et nous relions cette question à l'échantillon nécessaire de volumes de trafic. Toute la discussion est illustrée par un exemple.

Introduction

In a previous paper (Robillard 1973), I showed how the observed link volumes could be utilised for gaining information on the O-D matrix and network characteristics. The presentation was very general and left aside most of the operational aspects of the method which we now discuss. Most specifically, we want to study the estimate of the O-D matrix obtained using observed link volumes for the class of proportional assignment methods. These methods are essentially the ones which do not take into account explicitly or implicitly any constraints of capacities.

Counting the introduction, the paper is divided into six sections. The second section consists of the definitions and the notation required for the paper to be self-contained. The estimation problem is precisely defined and is shown to be equivalent to a general non-linear regression problem. The exact definition of the class of proportional assignment models is given in section three. For that class of models the estimation problem is shown to be equivalent to a linear regression problem which, unfortunately, turns out to be singular in most cases. However, everything is not lost ! Quite useful information concerning the O-D matrix can be extracted from the data. We characterise the class of linear functions of the O-D matrix entries that are estimable from the data. Moreover, in the following section we show that if the information on the cost of travel between every pair of origin and destination is available, a distribution method like the gravity model can be used to estimate the O-D matrix. We terminate this section by giving a necessary and sufficient condition on the set of observed links, such that the O-D matrix be estimable.

We mention briefly in the following section the practical implications of the hypothesis of normality on the residuals of the model. The last section deals with the corresponding non-linear model. The estimation of the parameters of this model can be obtained by using repeatedly a linear regression estimation procedure. All through the paper we use the same example to illustrate almost each statement.

Notation

A directed graph G is a pair (V,L) where V is an unordered set of elements called nodes and L is an unordered set of ordered pairs of nodes called arcs. The nodes are denoted by letters such as i, j, k and the arcs by pairs of letters such as (i,j), (i,k). We denote by O-D the set of ordered pairs of nodes which are the origin and destination nodes. We suppose for simplicity, that every origin communicates with every destination. T_p represents the number of trips which start at the origin i and terminate at the destination k for the pair $p = (i,k)$. The set of all T_p is denoted by T and is called the O-D matrix.

An assignment algorithm A evaluates for each arc (i,j) a traffic volume F_{ij}. The set of all F_{ij} is the assignment. The traffic volume F_{ij} is a function of the algorithm A, the network G, the set of demands T and a set of parameters θ which serve to quantify the network (lengths, speeds, etc.) We denote this dependence by writing

$$F_{ij} = H_A(G,\theta,T,(i,j))$$

We denote by f_{ij} the observed flow on the arc (i,j) and by E the set of arcs where the f_{ij} are available. (E is a subset of L.) We do not require the flow conservation equations to hold for the f_{ij}. We say that an assignment method reproduces faithfully the reality whenever the assignment produced by the method agrees with the observed traffic flows. If one adopts a measure of closeness which evaluates the similarity between the observed and the

assigned traffic, one could evaluate the relative quality of different assignment methods, or for a given method one could compare the relevance of several O-D matrices. Moreover, one could choose the O-D matrix which minimizes the measure of closeness as an estimate of the true O-D matrix. We suggest for this study the use of the measure $C(f,F)$ defined as

$$C(f,F) = \sum_{(i,j)\in E} (f_{ij}-F_{ij})^2$$

that is, the sum of square of the differences between the observed and calculated traffic flow volumes for the set of observed arcs. Through the F's C is a function of the algorithm A, the graph G, the demands T and the set of parameters θ.

Given G, A, E and θ an estimate of the true O-D matrix is the set $\hat{T}$ which minimizes the measure of closeness C. It can be viewed as a non-linear least square estimate of the O-D matrix. As presented this estimation procedure is extremely general. The interpretation of the estimates is not very apparent at this point. Many questions concerning the choice of the sampled arcs from which data are to be obtained, the reliability of the estimates and the fit of the model remain to be answered. For that purpose, we specialize the presentation to the case of proportional assignment models for which most of these questions are easily solved. Besides the important results obtained for these models, the discussion will be helpful in orienting our intuition for more complex models.

Proportional assignment models

For a given assignment model, a graph G and a set θ we denote by $F_{ij}^{k}(T_k)$ the flow assigned on the arc (i,j) when the O-D demands are null for all O-D pairs but the k-th for which it equals T_k. An assignment method is said to be proportional whenever the following properties hold for all G, θ and T:

$$F_{ij} = \sum_{k \in O\text{-}D} F_{ij}^{k}(T_k) \qquad \text{for all } (i,j) \text{ in } L$$

and

$$F_{ij}^{k}(T_k) = T_k F_{ij}^{k}(1) \qquad \text{for all } (i,j) \text{ in } L \text{ and } k \text{ in O-D}.$$

The first equation means that the total assigned flow on each arc equals the sum of the assigned flows obtained when the method is applied on each O-D pair separately. The name proportional for this class of methods is a consequence of the second equation. For example, if all the entries of the O-D matrix are doubled, the assigned flow will double the flow assigned with the original matrix. These two properties are common to the methods that do not consider implicitly or explicitly any constraints of capacities on the arcs.

The measure of closeness between the observed and the calculated flows can now be written as

$$C(f,F) = \sum_{(i,j) \in E} \left(f_{ij} - \sum_{k \in O\text{-}D} T_k F_{ij}^{k}(1)\right)^2 .$$

The estimation of the O-D matrix consists in the calculation of the $\hat{T}$'s which minimize the measure C . Equivalently, the $\hat{T}$'s can be viewed as the least square estimates of the linear regression model

$$f_{ij} = \sum_{k \in O\text{-}D} T_k F_{ij}^{k}(1) + e_{ij} \qquad (i,j) \in E$$

where e_{ij} is the residual term for the arc (i,j) . We assume that the residual terms are independently and identically distributed random variables with null mean and finite variance, i.e.

$$E(e_{ij}) = 0$$

$$\begin{aligned} \mathrm{Cov}(e_{ij} e_{kn}) &= 0 && (i,j) \neq (k,n) \\ &= \sigma^2 && (i,j) = (k,n) \ . \end{aligned}$$

Except for special results, the normality of the residuals is <u>not</u> required.

It is instructive at this point to look at a very simple example and to work out the details of the estimation procedure. We consider the graph represented in figure 1 with four arcs numbered 1 to 4 and five nodes, A B C D and X , two of them being, as indicated, origin nodes and two of them destination nodes.

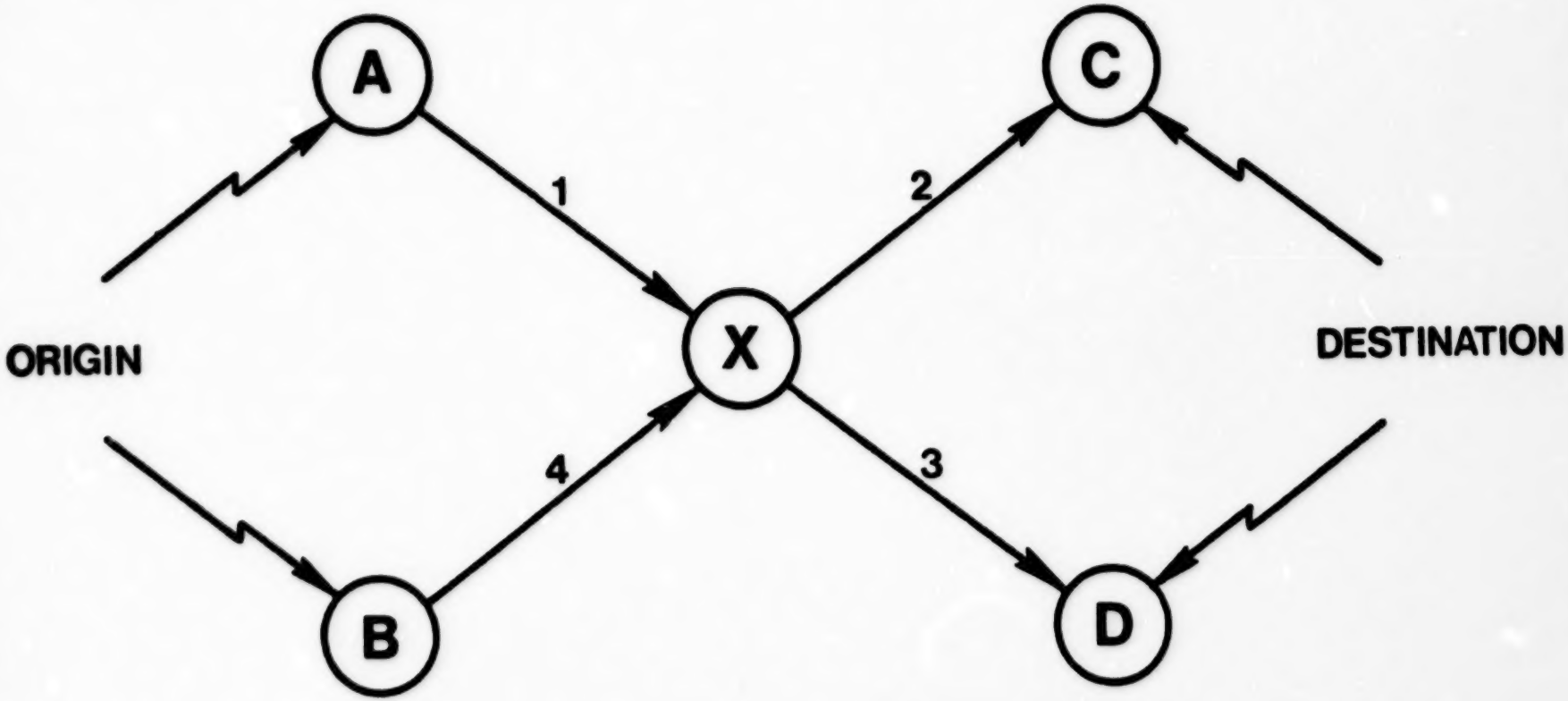

Figure 1

The set of O-D pairs consists of the pairs AC , AD , BC and BD . If one observation is available on each arc, the linear model can be written as

$$f_1 = T_{AC}F_1^{AC}(1) + T_{AD}F_1^{AD}(1) + e_1$$
$$f_2 = T_{AC}F_2^{AC}(1) + T_{BC}F_2^{BC}(1) + e_2$$
$$f_3 = T_{AD}F_3^{AD}(1) + T_{BD}F_3^{BD}(1) + e_3$$
$$f_4 = T_{BC}F_4^{BC}(1) + T_{BD}F_4^{BD}(1) + e_4$$

or in matrix notation as

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} = \begin{bmatrix} F_1^{AC}(1) & F_1^{AD}(1) & 0 & 0 \\ F_2^{AC}(1) & 0 & F_2^{BC}(1) & 0 \\ 0 & F_3^{AD}(1) & 0 & F_3^{BD}(1) \\ 0 & 0 & F_4^{BC}(1) & F_4^{BD}(1) \end{bmatrix} \begin{bmatrix} T_{AC} \\ T_{AD} \\ T_{BC} \\ T_{BD} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

and more briefly as

$$f = DT + e$$

where f is the vector of observations, D the design matrix, T the vector to be estimated and e the vector of residuals. From the definition, it is obvious that all the $F_{ij}^{k}(1)$ equal 1 for any assignment method. The linear model can now be rewritten as

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} T_{AC} \\ T_{AD} \\ T_{BC} \\ T_{BD} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

The important point to notice is the fact that the design matrix D is of rank 3 and is singular. As a consequence, the T's are not estimable, that is, the problem

$$\min_{T} \; C(f,F)$$

does not have a unique solution. Everything is not lost however !
It is well known that even if the T's are not estimable, certain linear combinations of them remain estimable (Rao 1973, chapter 4) and they are characterized as follows:

Lemma 1: $g'T$ is an estimable function if and only if g is a solution of the equation $(I-D'(D\,D')^{*}D)g = 0$ where $(D\,D')^{*}$ is the pseudo-inverse of the matrix $D\,D'$. (D' is the transpose of D) In other words, g must be contained in the linear vector space spanned by the columns of D'.

Continuing the example above

$$D\,D' = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix} \quad \text{and} \quad (D\,D')^{*} = 1/4 \begin{bmatrix} 4 & -2 & -2 & 0 \\ -2 & 3 & 1 & 0 \\ -2 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$I - D'(D\,D')^{*}D = 1/4 \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

From lemma 1 it is easily verified that $T_{AC} + T_{AD}$, the total number of trips originating from A, is estimable. Indeed,

$$T_{AC} + T_{AD} = g'T \quad \text{with} \quad g' = (1\ ,\ 1\ ,\ 0\ ,\ 0)$$

and

$$(I - D'(D\,D')^{*}g = 0\ .$$

Similar arguments show that the total number of trips starting at B and the total number of trips ending at C and D are also estimable, and their estimator can be obtained from the general result:

Lemma 2: if $g'T$ is an estimable function then the best linear unbiased estimator of $g'T$ is $g'\hat{T}$ where $\hat{T} = (D'D)^{*}D'f$. The variance of $g'\hat{T}$ is $g'(D'D)^{*}g\,\sigma^2$.

The unique estimator for $T_{AC} + T_{AD}$ is given by $\hat{T}_{AC} + \hat{T}_{AD}$ where

$$\begin{bmatrix} \hat{T}_{AC} \\ \hat{T}_{AD} \\ \hat{T}_{BC} \\ \hat{T}_{BD} \end{bmatrix} = 1/4 \begin{bmatrix} 2 & 2 & -2 & -2 \\ 1 & -1 & 3 & 1 \\ -1 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix}$$

thus $\hat{T}_{AC} + \hat{T}_{AD} = (3\,f_1 + f_2 + f_3 - f_4)$

and its variance equals $0.75\sigma^2$.

Similar formulae can be obtained for every estimable function of the T's .

To illustrate the theory with numbers, let us consider the same network with the observed flows as indicated in figure 2.

Notice that the flow conservation equations are not respected by the observed flows. The estimated flows at each of the origins and destinations are given by

$$\hat{T}_{AC} + \hat{T}_{AD} = 30.5$$
$$\hat{T}_{BC} + \hat{T}_{BD} = 10.5$$
$$\hat{T}_{AC} + \hat{T}_{BC} = 26.5$$
$$\hat{T}_{AD} + \hat{T}_{BD} = 14.5$$

The important fact is that even if we are unable to estimate the O-D matrix as such, we have enough information to get unique estimates for the total flows at each origin and destination. Moreover, the total number of trips that starts at the various origins equals the total number of trips that end at the various destinations. Part of the flow conservation law has been restored.

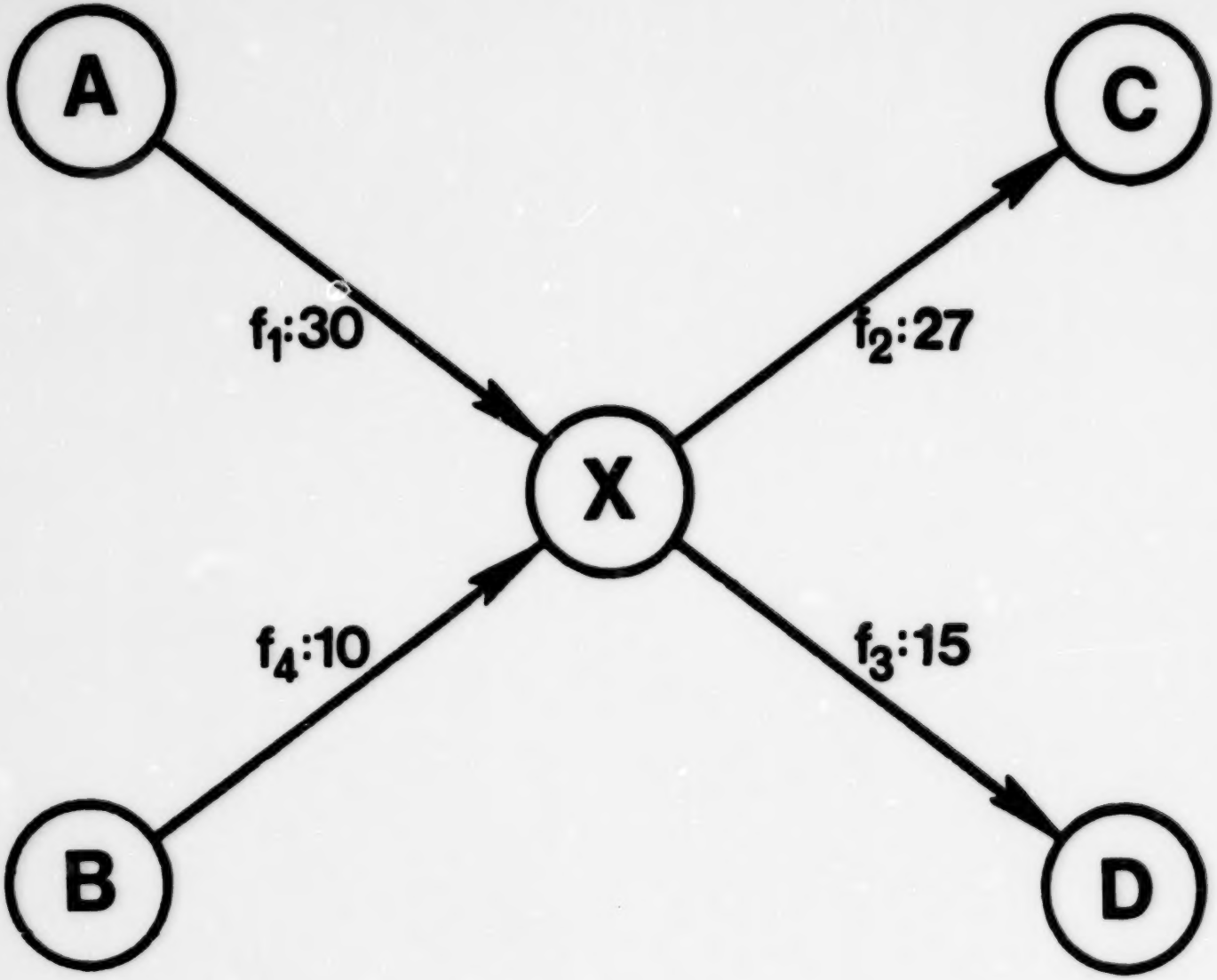

Figure 2

The distribution model

Suppose that for a given network, a given assignment model, and a given set of observed flows, one could estimate as in the last example, the trip volumes for each origin and each destination, then one has virtually solved the generation problem. From these results, and supplementary information, one could use a distribution model to evaluate the O-D matrix.

Continuing with our example, we might choose to structure the O-D flows using the generalized gravity model, that is,

$$
\begin{aligned}
T_{AC} &= G_{AC}\, a_A\, b_C \\
T_{AD} &= G_{AD}\, a_A\, b_D \\
T_{BC} &= G_{BC}\, a_B\, b_C \\
T_{BD} &= G_{BD}\, a_B\, b_D
\end{aligned}
$$

with the additional constraints

$$
\begin{aligned}
T_{AC} + T_{AD} &= \hat{T}_{AC} + \hat{T}_{AD} \\
T_{AC} + T_{BC} &= \hat{T}_{AC} + \hat{T}_{BC} \\
T_{BC} + T_{BD} &= \hat{T}_{BC} + \hat{T}_{BD} \\
T_{AD} + T_{BD} &= \hat{T}_{AD} + \hat{T}_{BD}
\end{aligned}
$$

The value G_{XY} measures the cost of travelling from the origin X to the destination Y. As shown by Hyman (1969) and Evans (1970) different choices

of these values give rise to different distribution models. If we suppose some values for the G's say $G_{AC} = 10$, $G_{AD} = 5$, $G_{BC} = 12$ and $G_{BD} = 10$, the unique solution in terms of T's of the above system is

$$T^*_{AC} = 20.65$$
$$T^*_{AD} = 9.85$$
$$T^*_{BC} = 5.85$$
$$T^*_{BC} = 4.65$$

In general, we can use this procedure provided that the values G are available and provided also that the total flows departing from an origin or ending at a destination are estimable. The last condition is a restriction on the choice of the sample, that is, on the design of the experiment.

Formally, let us denote by

$$W'_v T$$

the linear combinations of the T's representing the total flows departing (ending) at node v . If v is an origin and a destination node, two such linear combinations will be needed. Let W be the matrix of all the column vectors W'_v ; for our example W is written as

$$W = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

where the last column is the linear combination which corresponds to the total flow starting at node B . If there are n_1 nodes of origin, n_2 nodes of destination and n_3 pairs of origins and destinations possible, W is a $n_3 \times (n_1 + n_2)$ matrix of rank at most $n_1 + n_2 - 1$. From lemma 1 it is easily seen that the function $W_v'T$ is estimable if and only if the design matrix is such that

$$(I - D'(D\,D')^{*}W = 0$$

In other words

Lemma 3: A necessary and sufficient condition to be able to estimate the total traffic flows starting or ending at each node is to take a sample such that the linear space spanned by the column of the design matrix D' contains W .

As a corollary to the last lemma, we conclude that the sample size needed to be able to estimate the total traffic flows at each node is at least $n_1 + n_2 - 1$. One could take samples of greater size of course.

In many practical situations the data required for the estimation of the flows starting at an origin or ending at a destination are available. For example, on rapid transit systems turnstile counts may provide these data, perhaps by direction due to station layout; and traffic at major generators, (such as parking lots for factories or shopping centres)may provide similar data on auto flows.

If we want to qualify the precision of the estimators and the fit of the model, we need to hypothesize a statistical distribution of the residuals. It is often realistic to use the normal distribution for this purpose. Much information can now be deduced from the data (Draper and Smith,

1967). In particular, one could obtain confidence intervals for the estimated values, test the validity of the model with the data and qualify statistically the various parameters estimated from the data. The assumption of a known statistical distribution of the residuals brings a new dimension in the comprehension of the reliability of the results.

The companion non-linear model

The approach to estimate the O-D matrix described above can be summarized as a two step procedure:

Step 1: Using the linear model

$$f = DT + \varepsilon$$

estimate the total traffic flows departing and ending at each origin and destination node, that is, for each origin $m = 1,\ldots, M$ estimate the linear function $P_m = W_{1_m}{}'T$ and for each destination $n = 1,\ldots, N$ estimate the linear function $A_n = W_{2_n}{}'T$.

Step 2: Given the estimate $\hat{P}_m$ and $\hat{A}_n$ and the quantities G_{mn} which measure the relative cost of travelling from m to n, estimate T_{mn} the flow from m to n as the solution of the system

$$T_{mn} = R_m S_n G_{mn} \qquad m = 1, \ldots, M$$

$$\sum_m T_{mn} = \hat{A}_n \qquad n = 1, \ldots, N$$

$$\sum_n T_{mn} = \hat{P}_m$$

The last system has been shown to have a unique solution provided the A's and the P's and the G's are positive (Evans, 1970). It may happen in pathological cases that the estimates of the total flow volumes ending or departing at a node are negative.

To avoid this difficulty, one could use the non-linear model corresponding to the two step procedure described above, namely,

$$f_{ij} = \sum_{k=(m,n)} R_m S_n G_{mn} F_{ij}^k(1) + e_{ij}$$

The R's and the S's solution of this model can always be chosen positive. Let us notice the non uniqueness of the R's and S's estimates: indeed, one could multiply the R's by a constant and divide the S's by the same constant without changing the value of the f_{ij}. This multiplicity of solutions does not alter the value of the product of R's and S's.

In a recent paper, Lawton and Sylvestre (1971) have shown how to solve this non-linear system in a very efficient way. In summary, they notice that for given R's the model reduces to a linear regression model in the S's. They thus suggest to replace the S's by their linear regression solutions in terms of R's and solve for the R's only. This very simple procedure reduces the number of parameters to estimate and accelerate the convergence of the method. Let us illustrate the model by an example similar

to the one studied previously. Consider the network given in figure 3.

With $G_{AC} = 10$, $G_{AD} = 5$, $G_{AC} = 10$, $G_{BC} = 12$, the $F_{ij}^{k}(1)$ equals unity except for

$$F_5^{BD}(1) = 1 - F_4^{BD}(1) = 1 - F_3^{BD}(1) = p \ , \ o \leq p \leq 1 \ .$$

The value p depends on the assignment method used and on the relative costs of the two paths joining B and D . The non-linear regression model is now written as

$$\begin{aligned} 30 &= 10\ R_A S_C + 5\ R_A S_D \\ 27 &= 10\ R_A S_C + 12\ R_B S_C \\ 15 &= 5\ R_A S_D + 10\ p\ R_B S_D \\ 10 &= 12\ R_B S_C + 10\ p\ R_B S_D \\ 5 &= 10\ (1-p)\ R_B S_D \end{aligned}$$

The estimates of the T's were calculated for various values of p and are shown in figure 4 together with the corresponding residual sums of squares SSR . Notice the great variability of the measure SSR compared with the

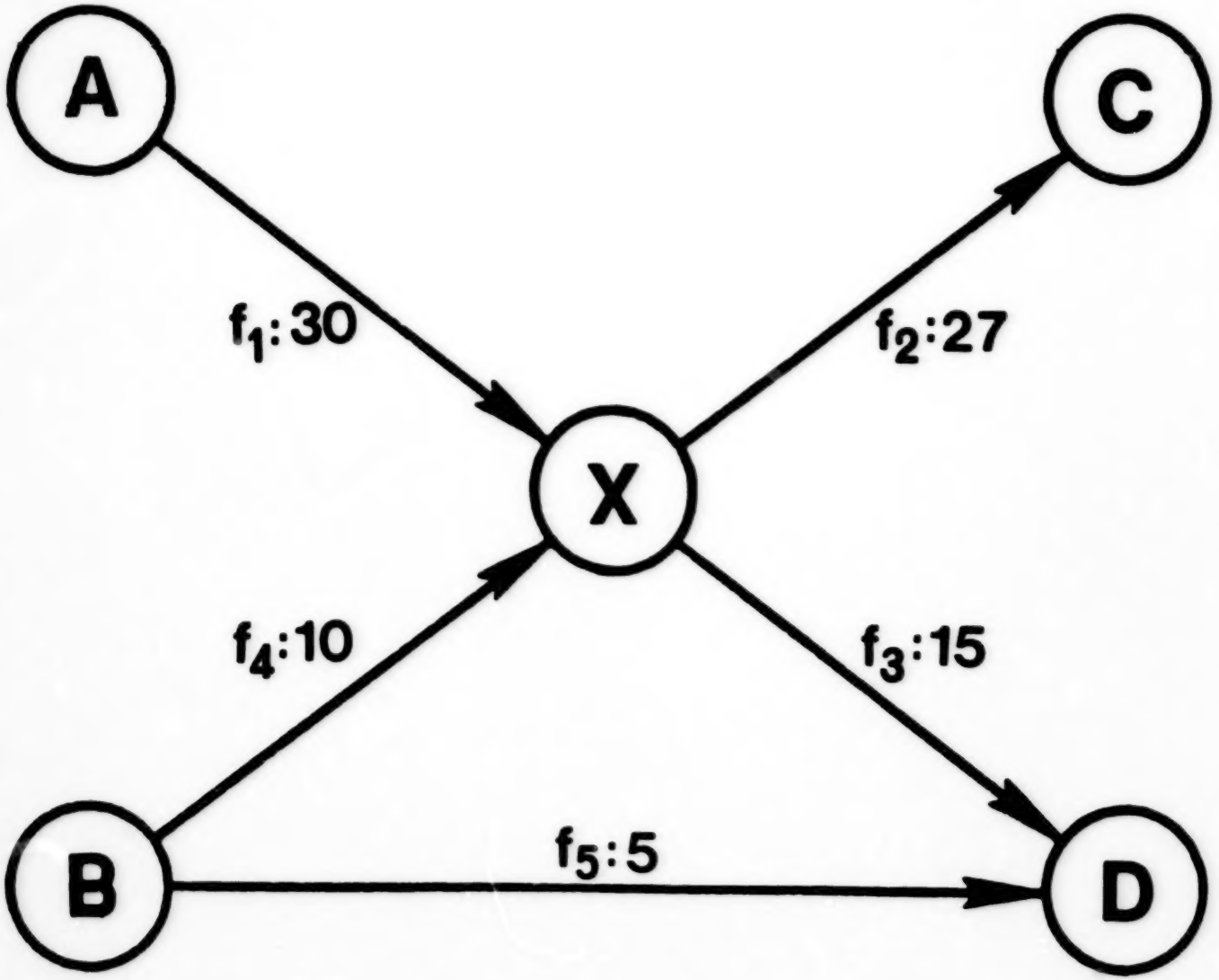

Figure 3

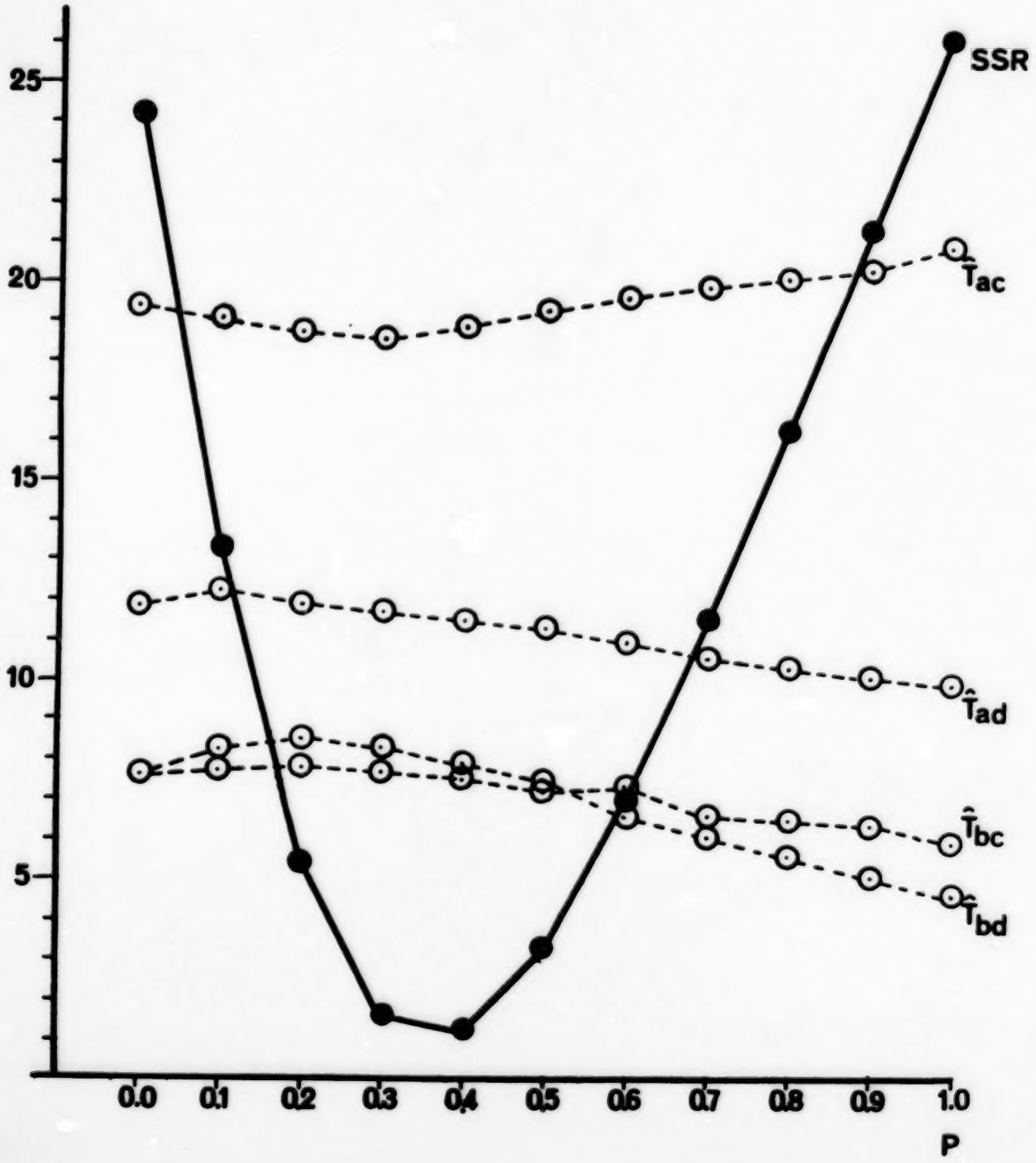
SSR
$\hat{T}_{ac}$
$\hat{T}_{ad}$
$\hat{T}_{bc}$
$\hat{T}_{bd}$
25
20
15
10
5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
P

Figure 4

relative stability of the estimated T's . This shows that an assignment method which allows the parameter p to take a value around 0.3 fits the data quite well. Indeed, one could estimate the p values as the ones for which SSR is minimum, that is, as $\hat{p} = 0.36$. The corresponding estimates of T's are $\hat{T}_{AC} = 18.86$, $\hat{T}_{AD} = 11.64$, $\hat{T}_{BC} = 7.64$, $\hat{T}_{BD} = 7.86$. This illustrates well the fact that the method of estimation proposed can be used to evaluate other quantities besides the O-D matrix. The problem of the choice of the sample cannot be handled here as simply as in the case of the linear regression model. We may suggest as a rule of thumb, to use the procedure for the linear model. The estimates obtained with the non-linear approach can be shown to be assymptotically normally distributed and only approximate confidence intervals can be evaluated following (Draper and Smith, 1967, chapter 10).

Conclusions and unsolved problems

The results discussed in this paper applies only to the class of proportional assignment models, which do not consider implicitly or explicitly any constraints of capacities on the arcs. This is a severe restriction which is imposed here in order to keep the mathematics tractable. The extension of the estimation method to non proportional assignment models and in particular to the case of equilibrium models (Dafermos and Sparrow, 1969) is important. The nature of such problem becomes more complex due to the non linearity of the model.

A second arbitrary decision made in the presentation of our method is the choice of the measure of closeness between the observed and the calculated flow which is the sum of square of residuals. Again this option was chosen mainly for mathematical tractability. It also inserts our method in the usual statistical least square estimation theory. The pitfalls of the least square theory are well known: our method is vulnerable in particular to the presence outlying observations and to the lack of normality of the data. We could have chosen instead a more "robust" measure of closeness such as the sum of absolute values of residuals. However the results on the estimability of the functions do not hold in such a case.

The purpose of this paper has been to study the theoretical aspect of estimating the O-D matrix from observed link volumes. We realize that a good deal more work is necessary before the results were applicable in practice. In particular we need to experiment with the method, study its possibilities and weaknesses and evaluate its utility while expending the theory.

References

Dafermos, S.C. and Sparrow, F.T. (1969), "The Traffic Assignment Problem for a General Network", Journal of Research, National Bureau of Standard, Vol. 37B, No. 2.

Draper, N.R., Smith, H. (1967), Applied Regression Analysis, John Wiley and Sons, N.Y.

Evans, A.W. (1970), "Some Properties of Trip Distribution Methods", Transpn. Res., Vol. 4, pp. 19-36.

Hyman, G.M. (1969), "The Calibration of Trip Distribution Models", Environment and Planning, Vol. 1, pp. 105-112.

Lawton, W.H., Sylvestre, E.A. (1971), "Elimination of Linear Parameters in Non-Linear Regression", Technometrics, Vol. 13, No. 3, pp. 461-467.

Rao, C.R. (1973), Linear Statistical Inference and Its Applications (second edition) John Wiley and Sons, N.Y.

Robillard, P. (1973), "Estimating the O-D Matrix and Network Characteristics from Observed Link Volumes". Proceedings of the International Conference on Transportation Research. First Conference - 1973, pp. 736-740.